

HARMONISED GUIDELINES FOR THE IN-HOUSE VALIDATION OF METHODS OF ANALYSIS (TECHNICAL REPORT)

Synopsis

ISO, IUPAC and AOAC INTERNATIONAL have co-operated to produce agreed protocols or guidelines on the “Design, Conduct and Interpretation of Method Performance Studies”¹ on the “Proficiency Testing of (Chemical) Analytical Laboratories”² on “Internal Quality Control in Analytical Chemistry Laboratories”³ and on the use of recovery information in analytical measurement⁴. The Working Group that produced these protocols/guidelines has been mandated by IUPAC to prepare guidelines on the In-House Validation of methods of analysis. These guidelines will outline minimum recommendations to laboratories producing analytical data on procedures that should be employed to ensure adequate validation of their methods before undertaking further validation involving other peer laboratories.

A draft of the guidelines will be discussed at an International Symposium on the Harmonisation of Quality Assurance Systems in Chemical Laboratory, the Proceedings from which will be published by the UK Royal Society of Chemistry.

DRAFT

CONTENTS

- 1 Introduction
 - 1.1 Background
 - 1.2 Approaches to In-House Method Validation
 - 1.3 Aspects of Method Performance Characteristics
 - 1.4 Measurement Uncertainty
 - 1.5 Existing In-House Method Validation Protocols and Measurement Uncertainty Standards/Guides
 2. Definitions and terminology used in guidelines
 3. The introduction of a new or unfamiliar method into the laboratory
 4. Selection of individual parameters to be assessed
 5. Statistical aspects of method validation
 - 5.1 The 'Ladder of Errors' and Other Sources of Error
 - 5.2 Higher Level Errors
 6. Uncertainty Estimation From Validation Studies
 - 6.1 Method Validation and the ISO *Guide*
 - 6.2 Method Validation
 - 6.3 Conduct of Validation Studies
 - 6.4 Relevance of prior studies
 - 6.5 Principles of Uncertainty Estimation Using Method Performance Data
 - 6.6 Uncertainty Estimation Using Prior Collaborative Method Development and Validation Study Data
 - 6.7 Uncertainty Estimation During In-House Development and Validation Studies
 - 6.8 Empirical Methods
 - 6.9 Ad-hoc Methods
 7. Conclusions
 8. Recommendations
 9. References
- Appendix I: Requirements For Individual Parameters By Which A Method May Be Characterised
- 1 Applicability
 - 2 Selectivity
 - 3 Calibration
 - 4 Trueness
 - 5 Precision
 - 6 Recovery
 - 7 Range
 - 8 Limit of Quantification
 - 9 Limit of Detection
 - 10 Sensitivity
 - 11 Ruggedness
 - 12 Fitness-for-Purpose

1 INTRODUCTION

1.1 Background

It is internationally recognised that a laboratory must take appropriate quality assurance measures to ensure that it is capable of and does provide data of the required quality. Such measures include:

- using validated methods of analysis,
- using internal quality control procedures,
- participating in proficiency testing schemes, and
- becoming accredited to an International Standard, normally the ISO/IEC Guide 25.

Aspects of the above have been previously addressed by the IUPAC Interdivisional Working Party on Harmonisation of Quality Assurance Schemes for Analytical Laboratories, specifically by preparing Protocols/Guidelines on method performance (collaborative) studies¹, internal quality control³ and proficiency testing².

Thus it may be seen that method validation is one, albeit essential, component of the measures that a laboratory should implement to allow it to produce reliable analytical data. There is a continuing need for reliable analytical methods for use in determining compliance with national regulations as well as international requirements in all areas of analysis. The reliability of a method is determined by some form of a validation procedure.

“Full” validation for an analytical method usually comprises an examination of the characteristics of the method in an inter-laboratory method performance study (*aka* collaborative study or collaborative trial). In some sectors, most notably food, the requirement for control analysts to use methods which have been fully validated is prescribed by legislation^{5, 6}. Internationally accepted protocols have been established for the “full” validation of a method of analysis by a collaborative trial, most notably the International Harmonised Protocol¹ and the ISO Procedure⁷. These protocols/standards require a minimum number of laboratories and test materials to be included in the collaborative trial to fully validate the analytical method. This usually requires a study design involving a minimum of 5 test materials, the participation of 8 laboratories reporting valid data, and most often includes blind replicates or split levels materials to assess within-laboratory repeatability parameters.

However, before a method is subjected to validation by a collaborative trial (i.e. to become fully validated) the method must undergo some validation within a single laboratory, usually in the laboratory which develops/modifies the analytical method. This validation may be regarded as “in-house method validation”.

As stated above, the ideal validated method is one that has progressed fully through a collaborative study in accordance with international harmonised protocols for the design, conduct and interpretation of method performance studies. It is not practical or necessary to require that all analytical methods used in the analytical laboratory be assessed at the ideal level. Limiting factors for completing ideal multi-laboratory validation studies include high costs, lack of sufficient expert laboratories available and willing to participate in such studies, and overall time constraints.

Thus “in-house” method validation should be undertaken to:

- ensure the validity of the method prior to the costly exercise of a formal collaborative trial,

- ensure that validated methods are being used correctly when used by analysts prior to undertaking a particular study, and
- provide evidence of the reliability of analytical methods if supporting trial data is not available or where the conduct of a formal collaborative trial is not feasible for either practical or economic reasons.

The need to permit methods to be assessed in-house without proceeding to a full validation through collaborative trial is becoming more common place. These Guidelines support that approach provided the laboratory carrying out the work is operating “under control”. The requirement for laboratories to so operate is either prescribed by legislation in some sectors of analysis, or is being demanded of them by their customers.

1.2 Approaches to In-House Method Validation

As indicated above the quality of an analytical result for laboratory depends on a number of functions, only one of which is the intrinsic quality of the analytical method. As laboratory quality standards have developed, and their need for attainment by a laboratory has become recognised, the necessity for a laboratory to use a fully validated method is reducing. In addition in a number of sectors it is increasingly difficult to fully validate by collaborative trial all analytical methods for practical and economic reasons. As a result there is now a move towards placing a greater emphasise on and use of methods of analysis which have only been validated “in-house” provided that the laboratory implements appropriate quality assurance procedures.

There are two significant approaches to in-house method validation, mainly the “traditional”, i.e. identifying and then evaluating certain set parameters, and the more recent “measurement uncertainty” approach. These approaches are both outlined in these guidelines.

1.3 Aspects of Method Performance Characteristics

Method validation may be described as the set of tests used to establish and document the performance characteristics of a method and thereby demonstrate that the method is fit for a particular analytical purpose and against which the method may be judged. The performance characteristics of the method are experimentally-derived values, for the fundamental parameters of importance in assessing the suitability of the method. The latter parameters include:

Applicability, selectivity, calibration, trueness, precision, recovery, range, limit of quantification, limit of detection, sensitivity, ruggedness and fitness-for-purpose

These are considered in these Guidelines, and more extensively in Appendix I.

1.4 Measurement Uncertainty

In 1993 ISO published its “Guide to the Expression of Uncertainty in Measurement”⁸, which was then interpreted by EURACHEM⁹. This recommended a then new approach to the evaluation of uncertainty of measurement which in chemical analysis may be taken as the uncertainty of any particular analytical result. There was a strong emphasis in that guide for measurement uncertainty to be evaluated using a “component-by-component” approach whereby the variances or

uncertainties within a method of analysis are identified and quantified as a standard deviation. These standard deviations are then combined to give an estimate of the overall standard deviation of the analytical method. These guidelines also comment on that approach for obtaining an estimate of the uncertainty of a method of analysis. The approach may be regarded as a logical development of the traditional approach but with several components being identified and quantified together.

1.5 Existing In-House Method Validation Protocols and Measurement Uncertainty Standards/Guides

A number of protocols/guidelines have been prepared on in-house method validation, most notably the following:

A Protocol on the Validation of Chemical Analytical Methods developed by the Nordic Committee on Food Analysis¹⁰,

A generic laboratory guide developed by EURACHEM produced by the UK Laboratory of the Government Chemist with the support of the UK Department of Trade and Industry Valid Analytical Measurement Initiative¹¹,

An Interlaboratory Analytical Method Validation Short Course developed by the AOAC INTERNATIONAL¹²,

A Guide to the Validation of Methods developed by the Dutch Inspectorate for Health Protection¹³,

A guide to Analytical Quality Assurance in Public Analyst Laboratories prepared by the UK Association of Public Analysts¹⁴,

A Guide to the Fitness for Purpose of Analytical Methods, prepared by a EURACHEM Working Group¹⁵, and

The subject was extensively discussed at a Joint FAO/IAEA Expert Consultation, December 1997, on the Validation of Analytical Methods for Food Controls, the Report of which is available¹⁶.

In addition “measurement uncertainty” has been extensively discussed, most notably in the:

ISO “Guide to the Expression of Uncertainty in Measurement”⁸,

The EURACHEM Guide to Quantifying Uncertainty in Analytical Measurement⁹, and

The NMKL Guide to the “Estimation and Expression of Measurement Uncertainty in Chemical Analysis”¹⁷.

These guidelines, prepared by the IUPAC Interdivisional Working Party on the Harmonisation of Quality Assurance Schemes for Analytical Laboratories, bring together essential elements of the above documents to give information which has been subjected to international acceptance and, more importantly, point the way forward with regard to in-house method validation.

2. DEFINITIONS AND TERMINOLOGY USED IN GUIDELINES

[To be completed]

3. THE INTRODUCTION OF A NEW OR UNFAMILIAR METHOD INTO THE LABORATORY

When a laboratory intends to use a method with which it's unfamiliar it is the responsibility of that laboratory to ensure that it is competent to use the method. The extent that laboratory has to undertake familiarisation with the method depends on the validation of the method and the competence of the laboratory. Some ideal suggestions as to the extent of validation and verification measures for different circumstances are given below:

The laboratory is to use a fully validated method

The method has been studied in a collaborative trial and so the laboratory has to verify that it is capable of achieving the published performance characteristics of the method (or is able to fulfil the requirements of the analytical task).

The laboratory is to use a fully validated method, but new matrix or new instruments is to be used

The method has been studied in a collaborative trial and so the laboratory has to verify that it is capable of achieving the published performance characteristics of the method (or is able to fulfil the requirements of the analytical task). In addition, the laboratory should extended the verification of trueness and precision and other appropriate characteristics, e.g. detection limit, in the new analytical system.

The laboratory is to use a well-established, but not collaboratively studied, method

The laboratory is to undertake verification, supplemented with a limited validation (e.g. with respect to reproducibility).

The method published in the scientific literature together with some analytical characteristics

The laboratory is to undertake verification, supplemented with limited validation (e.g. with respect to repeatability and reproducibility).

The method published in the scientific literature; no characteristics given

The laboratory is to arrange a full validation of the method and self-verification that it can achieve the determined characteristics.

The method has been developed in-house

The laboratory is to arrange a full validation of the method and self-verification that it can achieve the determined characteristics.

It is essential that appropriate validation or verification *is* always be carried out before taking a new method into routine use. Verification should be repeated in cases such as, for example, when:

major instruments are replaced,
new batches of major reagents, e.g. polyclonal antibodies, are taken into use,
changes made in the laboratory premises may influence the results,
methods are used for the first time by new staff, and
the laboratory takes an already validated and verified method into use after it has been out of use for a long time

4. SELECTION OF INDIVIDUAL PARAMETERS TO BE ASSESSED

There are individual parameters which may be used to characterise a method of analysis. However, when a method is to be characterised in-house it is important that the laboratory determines and agrees with its customer exactly which characteristics are to be evaluated. In a number of situations these may be laid down by legislation (e.g. the veterinary drug residues in food and pesticides in food sectors). The pesticide sector has a number of characteristics prescribed by legislation that have to be evaluated. The extent of the evaluation that a laboratory undertakes must meet the requirements of legislation.

A list of the possible parameters that may be considered and assessed is given in Section 1.3; a full description of the parameters is given in the Appendix.

5. STATISTICAL ASPECTS OF METHOD VALIDATION

5.1 The 'Ladder of Errors' and Other Sources of Error

Errors in analytical measurements arise from different sources and at different levels of organisation. In the past there has been a tendency for analysts or laboratories to neglect some of these sources and levels when constructing an uncertainty budget. This results in uncertainty estimates that are too narrow. There is a particular danger that such underestimates will occur in the context of in-house method validation.

There are many possible ways to represent the sources of error in the result of a measurement, but a particularly useful way is the following 'ladder of errors'[ref].

the result = the true value
+ the method bias
+ the laboratory bias
+ the run effect
+ the repeatability error.

In many (but not all) types of analysis the errors at each of these levels are typically of comparable magnitude, so cannot be safely ignored in validation.

The repeatability (within-run) error includes contributions from the familiar gravimetric and volumetric errors, calibration/evaluation random errors, heterogeneity of the test material, and variation in the efficacy of the chemical treatment stages of the analysis: some of these factors are addressed separately in validation. The run effect accounts for day-to-day variations in the analytical system, such as changes of analyst, batches of reagents, instruments, and the laboratory

environment (*e.g.*, temperature changes). In in-house validation, the run effect is studied by conducting a designed experiment with (say) duplicated analysis of an appropriate material in a number of separate runs. (Internal quality control also addresses the run effect, but because it is ongoing, is not included in the techniques of validation, which is for present purposes regarded as a one-off exercise.) Between-laboratory variation is clearly seen as a reality in the results of collaborative trials (method performance studies) [refs] and proficiency tests [refs], and between-method variation can sometimes be discerned in the results of the latter.

In addition to the ladder of errors, other types of validation must be considered when discussing the validation of a method. The validation of a method is limited to its application; *i.e.* a method is applied to a particular class of test material. If there is a substantial variation of matrix types within the defined class, there will be an additional source of variation due to within-class matrix effects. (Of course, if the method is subsequently used for materials outside the defined class (that is, outside the scope of the validation) an extra uncertainty of unknown magnitude is introduced.)

A further consideration is the fact that the magnitude of the uncertainty, from all sources and at all levels, is a function of concentration. This factor can be ignored only if the concentration range within which the method will be applied is narrow, covering (say) half of an order of magnitude. (This may well happen in manufacturing quality control, for example.) In other circumstances, it may be necessary to validate the method at several typical concentrations. In most instances, the measurement uncertainty increases absolutely with concentration, but often, at concentrations ranging well above the detection limit, the relative uncertainty approximates to a constant value. If measurements are to be restricted to that range, a single estimate of relative uncertainty may suffice for in-house validation. Such a practice should not be adopted uncritically as policy but rather tested in each instance.

5.2 Higher Level Errors

The potential fault that must be avoided in in-house method validation is the undue neglect of the higher rungs of the ladder. There are a few laboratories with special facilities where method bias and laboratory bias can be regarded as negligible, but that circumstance is wholly exceptional. (Note that if there is only one laboratory carrying out a particular analysis, then method bias and laboratory bias take on a different perspective). Normally the higher rungs have to be included in the uncertainty budget, but often they are more difficult to address than repeatability error and the run effect. In general to assess the respective uncertainties it is necessary to access to information gathered at higher levels in the ladder, *i.e.*, by 'top-down' methods. The most generally useful sources of top-down information are (i) statistics from collaborative trials (clearly not available in most situations of in-house method validation), (ii) statistics from proficiency tests and (iii) results from the analysis of certified reference materials.

Collaborative trials directly estimate the variance of between-laboratory biases. While there may be theoretical shortcomings in the design of such trials, there is no doubt that these variance estimates are appropriate for practical purposes. Consequently it is always instructive to test in-house validation by comparing the estimates of uncertainty with the results of collaborative trials. If the in-house result is substantially the smaller, it is likely that important sources of uncertainty have been neglected. (Alternatively, it may be that a particular laboratory wants to claim a smaller uncertainty than found in collaborative trials: such a laboratory would have to take extraordinary measures to justify such a claim.) If no collaborative trial has been carried out on the particular method/test material combination, a good estimate of the reproducibility standard deviation σ_H at an analyte concentration c can usually be obtained from the Horwitz function, $\sigma_H = 0.02c^{0.8495}$,

with both variables expressed as mass fractions [refs]. This information may be carried into the in-house area with minimum change.

Statistics from proficiency tests [refs] are particularly interesting because they provide information in general about the magnitude of high level biases combined and also, for the participant, information about total bias on specific occasions. The statistics (for example, robust standard deviations of the participants results for an analyte in a round of the test) can in principle be used in a way similar to reproducibility standard deviations from collaborative trials, *i.e.*, to obtain a benchmark for overall uncertainty for comparison with individual estimates from in-house validation. In practice, statistics from proficiency tests may be more difficult to access, because they are not systematically tabulated and published like collaborative trials, but only made available to participants. Of course, if such statistics are to be used they must refer to the appropriate matrix and concentration of the analyte. Individual participants in proficiency testing schemes can also gauge the validity of their estimated uncertainty by comparing their reported results with the assigned values of successive rounds. This, however, is an ongoing activity and therefore not strictly within the purview of in-house validation (which is a one-off event).

If a certified reference material [ref] of appropriate matrix is available, a within-house test allows a laboratory to assess laboratory bias and method bias in combination, by analysing the CRM a number of times in separate runs. The estimate of the combined bias is the difference between the mean result and the certified value. (Note that the mean of a number of runs is required: if the result of a single run were used, the run bias contribution might be atypical for the analytical system. Similarly, if the method bias alone (*i.e.*, separated from laboratory bias) is required, results from an interlaboratory study would have to be considered.) Appropriate certified reference materials are not always available, so other materials may perforce have to be used. Materials left over from proficiency tests sometimes serve this purpose and, although the materials may have unknown uncertainty, their use to estimate the magnitude of uncertainties of the upper rungs of the ladder is far better than having no information at all. A further alternative is to use spiking and recovery information [ref] to provide estimates of these biases, although there may be unmeasurable sources of uncertainty associated with these techniques.

Perhaps the least recognised aspect of uncertainty in validation is that due to matrix variation within the defined class of test material. The theoretical requirement for the estimation of this uncertainty component is for a representative collection of test materials to be analysed in a single run, their individual biases estimated, and the variance of these biases calculated. (Analysis in a single run means that higher level biases have no effect on the variance. If there is a wide concentration range involved, then allowance for the change in bias with concentration must be made.) If the representative materials are certified reference materials, the biases can be estimated directly as the differences between the results and the reference values, and the whole procedure is straightforward. In the more likely event that insufficient number of certified reference materials are available, recovery tests with a range of typical test materials may be resorted to, with due caution. Currently there is very little quantitative information about the magnitude of uncertainties from this source, although in some instances they are suspected of being large.

6. UNCERTAINTY ESTIMATION FROM VALIDATION STUDIES

6.1 Method Validation and the ISO Guide

In practice, the fitness for purpose of analytical methods applied for routine testing is most commonly assessed through method validation studies. Such studies produce data on overall

performance and on individual influence factors which can be applied to the estimation of uncertainty associated with the results of the method in normal use. This section provides guidance on the use of such data in estimating the overall uncertainty associated with a measurement result.

6.2 Method Validation

Method validation studies rely on the determination of overall method performance parameters. These are obtained during method development and interlaboratory study [refs], or following in-house validation protocols. Individual sources of error or uncertainty are typically investigated only when significant compared to the overall precision measures in use [ref]. The emphasis is primarily on identifying and removing (rather than correcting for) significant effects, leading to a situation in which the majority of potentially significant influence factors have been identified, checked for significance compared to overall precision, and shown to be negligible. Under these circumstances, the data available to analysts consists primarily of overall performance figures, together with evidence of insignificance of most effects and some measurements of any remaining significant effects. In following paragraphs, the collection of this data and its application in uncertainty estimation is discussed.

Validation studies for quantitative analytical methods typically determine some or all of the following parameters:

Precision

The principal precision measures include repeatability standard deviation s_r , reproducibility standard deviation s_R , (ISO 3534-1) and intermediate precision, sometimes denoted s_{zi} , with i denoting the number of factors varied (ISO 5725-3:1994)⁷. s_r indicates the variability observed within a laboratory, over a short time, using a single operator, item of equipment etc. s_r may be estimated within a laboratory or by inter-laboratory study. Interlaboratory reproducibility standard deviation s_R for a particular method may only be estimated directly by interlaboratory study; it shows the variability obtained when different laboratories analyse the same sample. Intermediate precision relates to the variation in results observed when one or more factors, such as time, equipment and operator, are varied within a laboratory; different figures are obtained depending on which factors are held constant. Intermediate precision estimates are most commonly determined within laboratories but may also be determined by interlaboratory study. The observed precision of an analytical procedure is an essential component of overall uncertainty, whether determined by combination of individual variances or by study of the complete method in operation.

Bias

The bias of an analytical method is usually determined by study of relevant reference materials or by spiking studies. Bias may be expressed as analytical recovery (value observed divided by value expected). Bias is expected to be negligible or otherwise accounted for, but the uncertainty associated with the determination of the bias remains an essential component of overall uncertainty.

Linearity

Linearity of response to an analyte is an important property where methods are used to quantify at a range of concentrations. Linear response to pure standards and to realistic samples may be determined. Linearity is not generally quantified, but is checked for by inspection or significance tests for non-linearity. Significant non-linearity is usually corrected for by non-linear calibration or eliminated by choice of more restricted operating range. Any remaining deviations from linearity

are normally sufficiently accounted for by overall precision estimates covering several concentrations, or within any uncertainties associated with calibration.

Detection limit

During method validation, the detection limit is normally determined only to establish the lower end of the practical operating range of a method. Though uncertainties near the detection limit may require careful consideration and special treatment, the detection limit, however determined, is not of direct relevance to uncertainty estimation.

Robustness or ruggedness

Many method development or validation protocols require that sensitivity to particular parameters be investigated directly. This is usually done by a preliminary 'ruggedness test', in which the effect of one or more parameter changes is observed. If significant (compared to the precision of the ruggedness test), either parameter variation is restricted further by prescription and the test repeated, or a more detailed study is carried out to measure the size of the effect, and a permitted operating interval chosen accordingly. Ruggedness test data can therefore provide information on the effect of important parameters which may not otherwise have been sufficiently accounted for.

Selectivity/specificity

Though loosely defined, both terms relate to the degree to which a method responds uniquely to the required analyte. Typical selectivity studies investigate the effects of likely interferences, usually by adding the potential interferent to both blank and fortified samples and observing the response. The results are normally used to demonstrate that the practical effects are not significant. However, since the studies measure changes in response directly, it is possible to use the data to estimate the uncertainty associated with potential interferences, given knowledge of the range of interferent concentrations.

6.3 Conduct of Validation Studies

The detailed design and execution of method validation studies is covered extensively elsewhere [refs] and will not be repeated here. However, the main principles as they affect the relevance of a study applied to uncertainty estimation are pertinent and are considered below:

Representativeness is essential. That is, studies should, as far as possible, be conducted to provide a realistic survey of the number and range of effects operating during normal use of the method, as well as covering the concentration ranges and sample types within the scope of the method. Where a factor has been representatively varied during the course of a precision experiment, for example, the effects of that factor appear directly in the observed variance and need no additional study unless further method optimisation is desirable.

In this context, *representative variation* means that an influence parameter must take a distribution of values appropriate to the uncertainty in the parameter in question. For continuous parameters, this may be a permitted range or stated uncertainty; for discontinuous factors such as sample matrix, this range corresponds to the variety of types permitted or encountered in normal use of the method. Note that representativeness extends not only to the range of values, but to their distribution.

In selecting factors for variation, it is important to ensure that the larger effects are varied where possible. For example, where day to day variation (perhaps arising from recalibration effects) is

substantial compared to repeatability, two determinations on each of five days will provide a better estimate of intermediate precision than five determinations on each of two days. Ten single determinations on separate days will be better still, subject to sufficient control, though this will provide no additional information on within-day repeatability.

It is generally simpler to treat data obtained from random selection than from systematic variation. For example, experiments performed at random times over a sufficient period will usually include representative ambient temperature effects, while experiments performed systematically at 24-hour intervals may be subject to bias due to regular ambient temperature variation during the working day. The former experiment needs only evaluate the overall standard deviation; in the latter, systematic variation of ambient temperature is required, followed by adjustment to allow for the actual distribution of temperatures. Random variation is, however, less efficient; a small number of systematic studies can quickly establish the size of an effect, whereas it will typically take well over 30 determinations to establish an uncertainty contribution to better than about 20% relative accuracy. Where possible, therefore, it is often preferable to investigate small numbers of major effects systematically.

Where factors are known or suspected to interact, it is important to ensure that the effect of interaction is accounted for. This may be achieved either by ensuring random selection from different levels of interacting parameters, or by careful systematic design to obtain both variance and covariance information.

In carrying out studies of overall bias, it is important that the reference materials and values are relevant to the materials under routine test.

Any study undertaken to investigate and test for the significance of an effect should have sufficient power to detect such effects before they become practically significant, that is, significant compared to the largest component of uncertainty.

6.4 Relevance of prior studies

Uncertainty estimates are almost invariably based at least partly on prior studies of method performance. It is accordingly necessary to demonstrate the validity of applying prior study results. Typically, such demonstration will consist of:

- Demonstration of comparable precision, typically by comparison of observed repeatability with published repeatability values where a previously validated method is implemented in a laboratory other than that performing the validation.
- Demonstrable control of bias, typically through determination of bias on relevant reference materials (see, for example, ISO Guide 33)[ref], by appropriate spiking studies, or satisfactory performance on relevant proficiency schemes or other laboratory intercomparisons
- Continued performance within statistical control as shown by regular QC sample results and the implementation of effective analytical quality assurance procedures.

Where the conditions above are met, and the method is operated within its scope and field of application, it is normally acceptable to apply the data from prior validation studies directly to uncertainty estimates in the laboratory in question.

6.5 Principles of Uncertainty Estimation Using Method Performance Data

The general approach to estimation of uncertainty from method performance data is that given in section 5.1. However, the aim is to provide the most economic estimate of overall uncertainty using overall method performance parameters. This is achieved by:

- **Reconciling information requirements with available data**

In identifying and quantifying the contributions to uncertainty, consider whether the available data accounts sufficiently for each source of uncertainty, whether by explicit study of the particular contribution or by implicit variation within the course of whole-method experiments. Identify any sources of uncertainty insufficiently accounted for by existing data, paying particular attention to the parameters varied during precision studies.

- **Obtaining further data as required**

Either obtain additional information from the literature or standing data (certificates, equipment specifications etc.), or plan experiments to obtain the required additional data. Additional experiments may take the form of specific studies of a single contribution to uncertainty, or the usual method performance studies conducted to ensure representative variation of one or more important factors.

The following sections provide guidance on the coverage and limitations of data acquired in particular circumstances and on the additional information required for an estimate of overall uncertainty.

6.6 Uncertainty Estimation Using Prior Collaborative Method Development and Validation Study Data

Although by definition collaborative trial data will not be available when validating a method in-house this section is included in these Guidelines for completeness. A collaborative study carried out, for example according to AOAC/IUPAC¹ or ISO 5725⁷ standards, to validate a published method, is a valuable source of data to support an uncertainty estimate. In particular, the reproducibility standard deviation s_R , serves as the basis for the standard uncertainty when the laboratory is “in control”.

Whether the estimate needs to be increased depends on the factors taken into account during the study. During the ‘reconciliation’ stage indicated above, the sources which need particular consideration are:

- **Sampling.** Studies rarely include a sampling step; if the method used in house involves sub-sampling, or the measurand (see Specification) is estimating a bulk property from a small sample, then the effects of sampling should be investigated and their effects included.
- **Pre-treatment.** In most studies, samples are homogenised, and may additionally be stabilised, before distribution. It may be necessary to investigate and add the effects of the particular pre-treatment procedures applied in-house.
- **Method bias.** Method bias is often examined prior to or during interlaboratory study, where possible by comparison with reference methods or materials. Where the bias itself, the

uncertainty in the reference values used, and the precision associated with the bias check, are all small compared to s_R , no additional allowance need be made for bias uncertainty. Otherwise, it will be necessary to make additional allowances.

- Variation in conditions. Laboratories participating in a study may tend towards the mean of allowed ranges of experimental conditions, resulting in an underestimate of the range of results possible within the method definition. Where such effects have been investigated and shown to be insignificant across their full permitted range, however, no further allowance is required.
- Changes in sample matrix. The uncertainty arising from matrix compositions or levels of interferences outside the range covered by the study will need to be considered.

For methods operating within their defined scope, and the factors above shown to be negligible or not relevant, the reproducibility standard deviation s_R , adjusted for concentration if necessary, may be used as the combined standard uncertainty.

Where additional factors apply, these should be evaluated in the form of standard uncertainties and combined with the reproducibility standard deviation in the usual way.

6.7 Uncertainty Estimation During In-House Development and Validation Studies

In-house development and validation studies consist chiefly of the determination of the method performance parameters. Uncertainty estimation from these parameters requires:

- The best available estimate of overall precision
- The best available estimate(s) of overall bias and its uncertainty
- Quantification of any uncertainties associated with effects incompletely accounted for in the above overall performance studies.

Precision study

The precision contribution should be estimated as far as possible over an extended time period, and chosen to allow natural variation of all factors affecting the result. Typical experiments include:

- Distribution of results for a typical sample analysed several times over a period of time, using different analysts and equipment where possible (A QC check sample may provide sufficient information)
- The distribution of replicate analyses performed on each of several samples.

NOTE: Replicates should be performed at materially different times to obtain estimates of intermediate precision; within-batch replication provides estimates of repeatability only.

- Formal multi-factor experimental designs, analysed by ANOVA to provide separate variance estimates for each factor.

Bias study

Overall bias is best estimated by repeated analysis of a relevant CRM, using the complete measurement procedure. Where this is done, and the bias found to be insignificant, the uncertainty associated with the bias is simply the combination of the uncertainty in the CRM value and the standard deviation associated with the bias check (adjusted for number of determinations).

NOTE: Bias estimated in this way combines bias in laboratory performance with any bias intrinsic to the method in use. Special considerations may apply where the method in use is standardised; see section

Where the reference material is only approximately representative of the test materials, additional factors should be considered, including (as appropriate)

- Effects of matrix or interferences which may differ from those in the reference material
- Differences in homogeneity; reference materials are frequently more homogeneous than test samples
- Any effects following from different concentrations of analyte; for example, it is not uncommon to find that extraction losses differ between high and low levels of analyte

Bias for a method under study is frequently determined against a reference method, by comparison of the results of the two methods applied to the same samples. In such circumstances, given that the bias is not statistically significant, the uncertainty is that for the reference method, combined with the uncertainty associated with the measured difference between methods. The latter contribution to uncertainty commonly appears as the standard deviation term used in the significance test applied to decide whether the difference is statistically significant.

EXAMPLE

A method (method 1) for determining the concentration of Selenium is compared with a reference method (method 2). The results from each method are as follows:

	\bar{x}	s	n
Method 1	5.40	1.47	5
Method 2	4.76	2.75	5

The standard deviations are pooled to give a pooled standard deviation s_c

$$s_c = \sqrt{\left(\frac{1.471^2 \times (5-1) + 2.750^2 \times (5-1)}{(5+5-2)} \right)} = 2.205$$

and a corresponding value of t:

$$t = \frac{(5.40 - 4.76)}{2.205 \sqrt{\left(\frac{1}{5} + \frac{1}{5}\right)}} = \frac{0.64}{1.4} = 0.46$$

t_{crit} is 2.3 for 8 degrees of freedom, so there is no significant difference between the means of the results given by the two methods. But the difference (0.64) is compared with a standard deviation term of 1.4 above. This value of 1.4 is the standard deviation associated with the difference, and accordingly represents the relevant contribution to uncertainty associated with the measured bias.

Overall bias is also commonly studied by addition of analyte to a previously studied material. The same considerations apply as for the study of reference materials (above). In addition, the differential behaviour of added material and material native to the sample should be considered and due allowance made. Such an allowance can be made on the basis of:

- studies of the distribution of error observed for a range of matrices and levels of added analyte
- comparison of result observed in a reference material with the recovery of added analyte in the same reference material
- judgement on the basis of specific materials with known extreme behaviour; for example, oyster tissue, a common marine tissue reference, is well known for a tendency to co-precipitate some elements with calcium salts on digestion, and may provide an estimate of 'worst case' recovery on which an uncertainty estimate can be based (e.g. by treating the worst case as an extreme of a rectangular or triangular distribution)
- judgement on the basis of prior experience

Bias may also be estimated by comparison of the particular method with a value determined by the method of standard additions, in which known quantities of the analyte are added to the test material, and the correct analyte concentration inferred by extrapolation. The uncertainty associated with the bias is then normally dominated by the uncertainties associated with the extrapolation, combined (where appropriate) with any significant contributions from the preparation and addition of stock solution.

NOTE: To be directly relevant, the additions should be made to the original sample, rather than a prepared extract.

It is a general requirement of the ISO Guide⁸ that corrections should be applied for all recognised and significant systematic effects.

Where the bias is not insignificant, but is nonetheless neglected for practical purposes, the uncertainty associated with bias should be increased by addition of a term equal to the measured bias.

NOTE The inclusion of a term equal to a measured bias is justifiable only to the extent that neglecting a significant bias is justifiable. It is better practice to report the bias and its uncertainty separately. Where this is not done, increasing the uncertainty estimate by including such a term simply avoids misleading any user of the reported result and uncertainty.

Additional factors

The effects of factors held constant during precision studies should be estimated separately, either by experimental variation or by prediction from established theory. Where practicable, the uncertainty associated with such factors should be estimated, recorded and combined with other contributions in the normal way.

Typical experiments include

- Study of the effect of a variation of a single parameter on the result. This is particularly appropriate in the case of continuous, controllable parameters, independent of other effects, such as time or temperature. The rate of change of the result with the change in the parameter can then be combined directly with the uncertainty in the parameter to obtain the relevant uncertainty contribution.

NOTE: The change in parameter should be sufficient to change the result substantially compared to the precision available in the study (e.g. by five times the standard deviation of replicate measurements)

- Robustness studies, systematically examining the significance of moderate changes in parameters. This is particularly appropriate for rapid identification of significant effects, and commonly used for method optimisation. The method can be applied in the case of discrete effects, such as change of matrix, or small equipment configuration changes, which have unpredictable effects on the result. Where a factor is found to be significant, it is normally necessary to investigate further. Where insignificant, the associated uncertainty is (at least for initial estimation) that associated with the robustness study.
- Systematic multifactor experimental designs intended to estimate factor effects and interactions.

Where the effect of an additional factor is demonstrated to be negligible compared to the precision of the study (i.e. statistically insignificant), it is recommended that an uncertainty contribution equal to the standard deviation associated with the relevant significance test be associated with that factor, at least until other contributions are shown to be substantially larger.

EXAMPLE

The effect of a permitted 1-hour extraction time variation is investigated by a t-test on five determinations each on the same sample, for the normal extraction time and a time reduced by 1 hour. The means and standard deviations were: Standard time: mean 1.8, standard deviation 0.21; alternate time: mean 1.7, standard deviation 0.17. A t-test uses the pooled variance of

$$\begin{aligned} & ((5-1) \times 0.21^2 + (5-1) \times 0.17^2) / ((5-1) + (5-1)) \\ & = 0.037 \end{aligned}$$

to obtain

$$t = (1.8 - 1.7) / \sqrt{0.037 \cdot (1/5 + 1/5)} = 0.82;$$

not significant compared to $t_{\text{crit}} = 2.3$.

But note that the difference (0.1) is compared with a calculated standard deviation term, of

$$\sqrt{0.037 \cdot (1/5 + 1/5)} = 0.3.$$

This is the contribution to uncertainty associated with the effect of permitted variation in extraction time.

Where an effect is detected and is statistically significant, but remains sufficiently small to neglect in practice, it is recommended that an uncertainty contribution equal to the measured effect combined with its statistical uncertainty be associated with the effect.

NOTE: See the note to section 5b.7.3

6.8 Empirical Methods

An 'empirical method' is a method agreed upon for the purposes of comparative measurement within a particular field of application where the measurand characteristically depends upon the method in use. The method accordingly defines the measurand. Examples include methods for leachable metals in ceramics and dietary fibre in foodstuffs.

Where such a method is in use within its defined field of application, the bias associated with the method is defined as zero. In such circumstances, bias estimation need relate only to the laboratory performance and should not additionally account for bias intrinsic to the method. This has the following implications.

Reference material investigations, whether to demonstrate negligible bias or to measure bias, should be conducted using reference materials certified using the particular method, or for which a value obtained with the particular method is available for comparison.

Where reference materials so certified are unavailable, overall control of bias is associated with the control of method parameters affecting the result; typically such factors as times, temperatures, masses, volumes etc. The uncertainty associated with these input factors must accordingly be assessed and either shown to be negligible or quantifiable with some assurance.

6.9 Ad-hoc Methods

Ad-hoc methods are methods established to carry out exploratory studies in the short term, or for a short run of test materials. Such methods are typically based on standard or well-established methods within the laboratory, but are adapted substantially (for example to study a different analyte) and will not generally justify formal validation studies for the particular material in question.

Since limited effort will be available to establish the relevant uncertainty contributions, it is necessary to rely largely on the known performance of related systems. Uncertainty estimation should accordingly be based on known performance on a related system or systems, combined with any specific study necessary to establish relevance of those studies. The following recommendations assume that such a related system is available and has been examined sufficiently to obtain a reliable uncertainty estimate.

As a minimum, it is essential that an estimate of overall bias and an indication of precision be available for the method in question. Bias will ideally be measured against a reference material, but will in practice more commonly be assessed from spike recovery. The considerations given previously then apply, except that spike recoveries should be compared with those observed on the related system to establish the relevance of the prior studies to the ad-hoc method in question. The overall bias observed for the ad-hoc method, on the materials under test, should be comparable to that observed for the related system, within the requirements of the study.

A minimum precision experiment consists of a duplicate analysis. It is, however, recommended that as many replicates as practical are performed. The precision should be compared with that for the related system; the standard deviation for the ad-hoc method should be comparable.

NOTE: It is recommended that the comparison be based on inspection; statistical significance tests (e.g. an F-test) will generally be unreliable with small numbers of replicates and will tend to lead to the conclusion that there is 'no significant difference' simply because of the low power of the test.

Where the above conditions are met unequivocally, the uncertainty estimate for the related system may be applied directly to results obtained by the ad-hoc method, making any adjustments appropriate for concentration dependence and other known factors.

Where these conditions are not met, it is not recommended that an uncertainty estimate be provided. Where an indication of reliability is nonetheless required, it is suggested that the measured bias and 95% confidence interval of the replicated results are reported, with the caveat that systematic effects on the result were not fully investigated.

7. CONCLUSIONS

It is now internationally recognised by both customers and producers of analytical laboratory information that that information be produced to an acceptable quality level. The laboratory undertaking the analysis must always be "in control" and use an appropriate analytical method. Ideally, the laboratory should agree with its customer what "quality" of method of analysis it should use. If a method which has been collaboratively tested is available, the laboratory will have to demonstrate why it is not used (typically that the method is dated, exhibits unacceptable performance characteristics when applied to the task in hand etc.). However, there are many situations when it is not possible or practicable to use a collaboratively tested method. There are a number of sectors of analytical chemistry where the concept of the "fully validated" method of analysis is not used or within a sector is not practicable because of the expense required to validate methods of analysis for all available analyte/matrix combinations. Thus, the need arises for a laboratory to use a method which has been validated in-house but which can be shown to possess certain analytical characteristics.

As a result of the above it is now internationally recognised that there is a need for guideline information to be prepared on In-House method validation. The need arises because of a number of different reasons, notably where the full validation of a method of analysis cannot be carried out for practical or economic reasons or where a new method is to be introduced into a laboratory.

8. RECOMMENDATIONS

The following recommendations are made regarding the use of In-House Method Validation:

- Wherever possible and practical a laboratory should use a method of analysis which has had its performance characteristics evaluated through a collaborative trial conforming to an international protocol.
- There are many situations where such methods are not available and will not be developed. In these situations a method must be validated in-house before being used to generate analytical data for a customer. The laboratory and its customer must identify which characteristics are to be determined during the in-house validation process.
- In-house validation requires the laboratory to select appropriate characteristics from the following for evaluation: applicability, selectivity, calibration, accuracy, precision, range, limit of quantification, limit of detection, sensitivity, ruggedness and practicability.
- Evidence that these characteristics have been assessed must be made available to customers of the laboratory if required by the customer.
- Evidence that these characteristics have been assessed must be made available to customers of the laboratory if required by the customer.
- In-house validated methods may then be used provided by a laboratory provided the laboratory can demonstrate that it is “in control”, i.e. is using appropriate internal quality control procedures, is participating in proficiency testing schemes and, ideally, is third party assessed for competency.

9. REFERENCES

1. "Protocol for the Design, Conduct and Interpretation of Method Performance Studies", W Horwitz, *Pure Appl. Chem.*, 1988, **60**, 855- 864, revised W. Horwitz, *Pure Appl. Chem.*, 1995, **67**, 331-343.
2. "The International Harmonised Protocol for the Proficiency Testing of (Chemical) Analytical Laboratories", M Thompson and R Wood, *Pure Appl. Chem.*, 1993, **65**, 2123-2144. (Also published in *J. AOAC International*, 1993, **76**, 926-940.
3. "Harmonised Guidelines For Internal Quality Control in Analytical Chemistry Laboratories", Michael Thompson and Roger Wood, *J. Pure & Applied Chemistry*, 1995, **67(4)**, 49-56.
4. "Harmonised Guidelines for the Use of Recovery Information in Analytical Measurement", Michael Thompson, Steven Ellison, Ales Fajgelj, Paul Willetts and Roger Wood, *J. Pure & Applied Chemistry*, 1999, **71(2)**, 337-348.
5. "Council Directive 93/99/EEC on the Subject of Additional Measures Concerning the Official Control of Foodstuffs", O. J., 1993, L290.
6. "Procedural Manual of the Codex Alimentarius Commission, 10th Edition", FAO, Rome, 1997.
7. "Precision of Test Methods", Geneva, 1994, ISO 5725, Previous editions were issued in 1981 and 1986.
8. "Guide to the Expression of Uncertainty in Measurement", ISO, Geneva, 1993.
9. "Quantifying Uncertainty in Analytical Measurement", EURACHEM Secretariat, Laboratory of the Government Chemist, Teddington, UK, 1995, EURACHEM Guide (under revision).
10. "Validation of Chemical Analytical Methods", NMKL Secretariat, Finland, 1996, NMKL Procedure No. 4.
11. "Method Validation - A Laboratory Guide", EURACHEM Secretariat, Laboratory of the Government Chemist, Teddington, UK, 1996.
12. "An Interlaboratory Analytical Method Validation Short Course developed by the AOAC INTERNATIONAL", AOAC INTERNATIONAL, Gaithersburg, Maryland, USA, 1996.
13. "Validation of Methods", Inspectorate for Health Protection, Rijswijk, The Netherlands, Report 95-001.
14. "A Protocol for Analytical Quality Assurance in Public Analysts' Laboratories", Association of Public Analysts, 342 Coleford Road, Sheffield S9 5PH, UK, 1986.
15. "The Fitness for Purpose of Analytical Methods", EURACHEM Secretariat, Internet Version, 1998
16. "Validation of Analytical Methods for Food Control", Report of a Joint FAO/IAEA Expert Consultation, December 1997, FAO Food and Nutrition Paper No. 68, FAO, Rome, 1998

17. “*Estimation and Expression of Measurement Uncertainty in Chemical Analysis*”, NMKL Secretariat, Finland, 1997, NMKL Procedure No. 5.

DRAFT

APPENDIX I: REQUIREMENTS FOR INDIVIDUAL PARAMETERS BY WHICH A METHOD MAY BE CHARACTERISED

The requirements for the individual parameters by which a method may be characterised are given as follows:

1 Applicability

The scope of application, or applicability, of the method should identify the matrix, analyte or species being measured, its concentration range and the type of study/monitoring effort for which the procedure, as judged from its performance characteristics, is suited. It should also describe the known limitations of the method.

Analysts should ensure that scope of the method chosen is suitable in terms of the matrix, analyte or species being measured, its concentration range and the type of study/monitoring effort described.

2 Selectivity

The selectivity is the ability to discriminate between the analyte or species to be determined and other materials in the test sample.

The extent to which the method responds only to the analyte, species or component which is intended to be measured should be evaluated. If the response is not specific to this measurand then it is a requirement that the method responds primarily to the analyte of interest and is little affected by other substances. Arrangements should therefore be made to correct, as far as possible, this response when it is affected by other substances or form(s) of the analyte present. Situations where the response of the analyte is affected include when:

- 1) The matrix and/or other analytes influence, by enhancing or suppressing, the sensitivity of the analytical measurement (the matrix effect).
- 2) Some components in the matrix contribute to the response of the analytical measurement, without influencing its sensitivity for the component of interest (the interference effect).
- 3) There is a combination of a matrix effect and an interference effect.

It may not be possible to attain or demonstrate specificity in all cases, either because known interferences remain or because of the practical limitation of testing all substances which may potentially have such an interfering effect.

If matrix blanks are available, analyse a minimum of 3 matrix blanks, 3 method blanks and the lowest concentration standard.

If matrix blanks are not available, analyse a minimum of 3 positive samples, 3 method blanks and a standard that corresponds to the analyte concentration in the positive sample.

Assess the magnitude and variability of the background signals from these blanks and the significance of any interferences or contamination relative to the lowest analyte level specified in the scope of the method. Ensure the absence of unacceptably high background signals from the blank by optimising reagent/equipment choice, analyte extraction, clean-up, and/or chromatographic separation efficiencies to eliminate the interference(s). Typically responses from interfering substances should be <1% of the response of the lowest concentration measured.

For the positive or spiked sample and the calibration standard, information in support of the qualitative selectivity of the method may be obtained by the use of alternative separation methods or detection techniques to confirm the identity and purity of the analytical signal. Supporting qualitative data can be provided by spectral data (UV, mass, IR, atomic, emission, etc.) or elemental analysis, to confirm whether the analyte of interest is the only species being measured. Co-chromatography, in conjunction with peak shape/peak purity analysis, may in addition be carried out for chromatographic techniques.

If the detection technique of the method being examined does not itself provide qualitative data (e.g. FID or refractive index detectors), then the additional use of an alternative detection approach, to provide information on the species being measured, is essential. However relying on the closeness of agreement of data between techniques, is not recommended for demonstrating selectivity unless one of the additional methods used is a standard method with known/demonstrated selectivity for the analyte in the matrix being studied.

3 Calibration

The calibration or standard curve is a graphic representation of the measuring signal (the response variable) as a function of the quantity of analyte or measurand.

The calibration curve, or response profile, relating the magnitude of the measuring signal to the amount or concentration of the analyte species, should be determined over an appropriate range of analyte concentrations, using calibration standards presented in a suitable solvent or sample matrix. A graph should be plotted of the measuring signal on the y-axis against the known concentration, or a related parameter, on the x-axis, to assess whether the curve exhibits a linear relationship or an alternative form, for example quadratic or cubic. The correct form of the calibration plot, and the mathematical formula which describes it, should be identified.

A sufficient number of standard solutions is needed to define the response profile in relation to concentration, the number required being a function of the range of concentration and the form of the curve. Determinations should be made on reference samples or on blank samples to which the analyte has been added. The response in solvent and matrix should be compared. The calibration is valid in the case of a standard prepared in solvent only if the analyte response from the calibration standard is the same as an equal amount of the analyte in the sample extract.

The calibration standards should be run in random order, the calibration standards being measured under the same conditions as those subsequently to be used for the test samples. The random error associated with the calibration can be minimised by replication of the calibration measurements. The spacing of the concentration levels and the replication pattern depends on the precision required over the range studied. Evenly distributed concentrations are often appropriate. The range of concentrations should cover at least 0-150% of the expected test result or specification, for the sample to be analysed. It should be noted that the factors which affect the magnitude of the

instrument response to a given concentration are many and complex and often vary considerably from day to day.

Once the calibration graph has been established, the analyte concentrations in test materials can be determined by interpolation from the graph. Extrapolation of the curve above or below the concentrations tested is not acceptable except at 'zero' and for the method of standard additions.

When a linear range exists within a wider non-linear response relationship between analyte concentration and resulting measurement response, it may be convenient to choose to work within the linear region. A minimum of 5 standard concentrations (excluding zero) is typically required to prepare a linear calibration curve although, providing that the requirements regarding precision and trueness are satisfied in the linear range, a reduced number of standards (i.e. <5) may be used for calibration purposes. If the calibration function does not change over the time-period of the study, then it may allow one standard to be used for quantitation. It is first necessary, however to establish the linearity of response. A number of procedures may be used to assess proportionality and test for linearity. Simple approaches include plotting the response factors (area/concentration) versus the concentrations or plotting the individual residuals against the respective values of y on the fitted straight line. More recently tests for non-linearity based on the analysis of the residual variance from a regression have been described which allow 'lack of fit' to be distinguished from pure error. The correlation coefficient (r) or the coefficient of determination (r^2) should not be used to determine the appropriateness of the linear model, but to describe it. In addition to quoting the correlation coefficient, the description of the linear calibration function is not complete without a statement of the confidence limits of the slope and y-intercept of the regression line. When carrying out the regression analysis, all data points, not averaged data, should be used.

For non-linear curves more standard solutions (typically >5) are required, the standard curve in the relevant concentration range being based on sufficient points to accurately determine the non-linear response function.

In the linear case, non-weighted regression should be used when the variance of the response variables is constant across the calibration range (homoscedasticity), however, a good assessment of this model requires at least 5 or 6 concentration levels and multiple measurements at each concentration level. The residuals plot is a valuable tool for assessing the appropriateness of the model for the data set.

When the assumption of a constant variance for the response variables across the calibration range, is violated (heteroscedasticity), weighted least squares regression will provide a more accurate assessment of the slope, intercept and confidence intervals for the regression line. A good estimate of the variance at each concentration is required otherwise the weighted regression may be more biased than the unweighted regression.

The use of only one standard solution as a reference for calculating the concentration of an unknown assumes that the y-intercept is zero within error ($y = mx$) and the change in response is directly proportional to the change in analyte concentration. The approach may be regarded as valid if the confidence intervals for the regression line are acceptably narrow relative to the precision required for the assay. There is less of a limitation if the concentration in the standard solution is chosen so that it is approximately equal to the expected concentration in the sample solution. The possibility of trends in the response factors should be investigated. A single point calibration may not be justified, and a measurement bias will occur if this approach is used inappropriately.

Multi-variate linear regression is used when there are 2 or more independent variables affecting a dependent variable.

4 Trueness

Trueness is the closeness of agreement between a test result and the accepted reference or true value of the property being measured, for example the true content of a specific analyte in a sample.

The accuracy or trueness of an analytical procedure should be evaluated, in terms of bias, through the analysis of either reference materials or spiked samples or by comparison with an alternative method. The approach which is chosen depends on the intended use of the method being validated and the resources available. The following approaches for determining accuracy are listed in order of desirability in terms of providing increased measurement reliability:

- the use of certified reference material;
- the use of a traceable reference material or material prepared 'in-house';
- the use of a reference method/standard method with little or no systematic error;
- the use of the method when participating in a proficiency testing scheme,
- the use of spiked samples, based on blank or positive samples.

Preference must be given to the use of a certified reference material (CRM) over the use of other reference materials, or a material prepared in the laboratory (in-house material) containing a known amount of the analyte. Whether a reference material or an in-house control material is used, in each case the material should be matched, as far as possible, in terms of analyte concentration and matrix to the test sample being analysed. The content of the analyte in in-house control materials must be thoroughly investigated, preferably using two or more analytical methods based on different physical chemical principles and, if possible, based on determinations carried out in different laboratories. An in-house material or non-certified reference should, whenever possible, be calibrated against a certified reference material. Caution should be exercised in the use of CRMs and other control materials since they may, on occasion, be more homogenous than ordinary food samples or not go through the same homogenisation steps as the unknown samples. The analysis of certified reference materials does not alone verify the accuracy of an analytical method. The analysis of reference materials should be supplemented with other quality criteria, such as for example recovery tests (spiking).

When a reference method is available, the accuracy of the test method can be examined by analysing the same samples with the method to be verified or validated and the reference method. If the reference method is not in routine use in the laboratory, it is not justified to introduce the method only in order to evaluate a new method. In such cases it is recommended that samples are sent to a laboratory having the necessary competence regarding the reference method, preferably an accredited laboratory.

The accuracy of the analytical method may also be assessed by using the method when participating in a proficiency test, devised to examine samples corresponding to those for which the candidate method is intended. It should be noted that the documented accuracy holds only for relevant analyte levels and matrices.

If a reference material is not available and there are no other methods appropriate for comparison, then accuracy is investigated by spiking, either into a representative matrix blank or into a sample containing a low level of analyte (positive sample). By spiking a suitable known amount of the analyte, or a chemical containing the analyte, into a test portion of a sample having a known concentration of the analyte, the recovery of the analyte can be determined. This is achieved by analysing the spiked test portion along with the original sample with no added analyte.

The recovery (%) of the added analyte is calculated as 100 x the difference in averages divided by the amount of the added amount:

$$[(\text{Analysed amount} - \text{Original amount in the sample})/(\text{Added amount})] \times 100$$

The added amount should correspond to the level normally present in the sample material. Such recovery tests offer a limited control of the systematic error by checking the recovery of the analysis. The technique is especially useful in connection with unstable analytes, or when only a limited number of determinations are to be made. It is recommended that recovery tests are carried out in the relevant concentration ranges. If the method is used on several levels, recovery tests should be carried out on at least two levels. It should be borne in mind, when using spike recovery data, that the analyte is not necessarily introduced into the sample matrix in the same manner as the analyte is incorporated into the actual samples. This means that there may be a difference in chemical form between the analyte in the authentic sample and in the added synthetic compound. Guidance on the conduct of spiking/recovery experiments has been given elsewhere.

As a minimum, when carrying out the spiking/recovery experiments:

- the spiking procedure must be precise and accurate

- the spike should be added prior to any sample extraction procedures

- the method of fortification must be described

- the amount of analyte spiked should not be less than the amount estimated to be present in the sample, the background level being subtracted to give % recovery

- when spiking positive samples, qualitative data should be obtained to demonstrate that only the analyte of interest is being measured.

- depending on the variability of the background analyte signal and the spike level it may be necessary to prepare many replicates to get a good estimate of the average analyte recovery and the uncertainty in this value. A t-test may be used in order to demonstrate whether or not the obtained recoveries are significantly different from 100%.

In all of the above cases, when using reference materials, spiked samples or comparison with an alternative method, bias should be assessed across the concentration range of interest and for each type of matrix. In addition, it is important that the studies conducted should also assess the variability of the bias, within-run and between run. To do this, it is generally recognised that at least 7 (preferably ≥ 10) test measurements, at each concentration level need to be carried out for acceptable measurement reliability in most cases. It generally holds that more determinations should be made the lower the analyte concentration is, because random errors increase with decreasing concentrations. It is recommended that for residue method validations, the variability of

bias is assessed at 2 representative levels, as a minimum, which cover the working range of the method. At each of the concentration levels a minimum of 5 replicates should be analysed in each of 4 batches. It is considered reasonable to analyse one batch of samples every 1-2 days.

Although the assessment of the true content of an analyte in a sample will always have an uncertainty associated with it, the absolute value of bias can be determined ideally by traceability of the measurand to the true value of an internationally accepted standard.

5 Precision

[Need to include Horwitz Curve as a fitness-for-purpose criterion]

Precision is the closeness of agreement between independent test results obtained under stipulated conditions.

The precision of the method should be characterised by the repeatability, and the internal reproducibility, which is a measure of intermediate precision obtained over a longer time period within the same laboratory. Repeatability is assessed in terms of within-batch or within-run precision data, while for internal reproducibility, between-batch or between-run precision data should be quoted, as described in ISO 5725 part 3⁵. It is important to make clear in the case of internal reproducibility, however, whether this relates to data collected over an extended time period only or whether it also covers different batches of reagent, different calibrations, different operators and/or the use of different equipment.

Precision characteristics may be determined using reference materials, control materials or actual analytical samples, but synthetic solutions which do not represent the analytical test materials should not be used. With the exception of the latter case - the use of actual analytical samples - relevant information may already be available from the studies carried out to assess the variability of the bias. Notwithstanding this, it is desirable to have measures of the precision characteristics of the method when applied to actual analytical samples which have been taken through the whole analytical procedure. For assessing the repeatability and internal reproducibility in the case of actual analytical samples, as when assessing the variability of bias, it is generally recognised that at least 7 (preferably ≥ 10) test measurements need to be carried out at each concentration level for acceptable measurement reliability, with relatively more determinations being required at lower analyte concentrations. For residue method validations, the precision characteristics should be assessed at a minimum of 2 representative levels, to cover the working range of the method.

It is recommended that one of the following procedures is adopted:

- i) At each of the concentration levels a minimum of 5 replicates should be analysed in each of 4 batches. It is considered reasonable to analyse one batch of samples every 1-2 days. The use of ANOVA statistics provides estimates of the repeatability and internal reproducibility for these samples and confidence intervals may be calculated for the standard deviations quoted in each case.
- ii) The repeatability and internal reproducibility may be assessed independently by, in each of the conditions, making at least ten determinations from the same sample material. Calculate the standard deviation either from results of replicate single determinations or from duplicate determinations. The standard deviation of single determinations is an estimate of the

distribution around the average, whereas the standard deviation derived from duplicate determination is an estimate of the variation between two single determinations.

In cases when only an initial indication of the internal reproducibility or within-laboratory precision is required, this may best be derived by carrying out replicate determinations at one or more appropriate levels of concentration. Not less than 6 pairs of replicate analytical results should generally be used. The appropriate level of concentration is dependent on a number of factors but generally will be the level of greatest interest, e.g. a statutory maximum level.

It is important that it is exactly documented how, and on what materials repeatability and internal reproducibility have been estimated (e.g. reference materials, control materials, synthetic solutions or actual analytical samples). When recording values of precision characteristics, the number of replicate determinations carried out should be stated.

It should be kept in mind that the measures of precision are generally dependent on the concentration of the analyte and the analytical technique being employed, so that in cases where the method under study is intended to be used over a large concentration range, the precision characteristics should be estimated at several concentration levels across the working range, for example, at low, medium and high levels. It is recommended that the estimation of precision is repeated in whole or in part during validation work. If the between-run standard deviation is unacceptably larger than the within-run standard deviation, then the reason for the added variability should be investigated, identified and controlled if possible.

For qualitative analyses, precision cannot be stated as a standard deviation, since this type of analysis in practice is a yes/no measurement at a given threshold concentration of the analyte. In this situation the precision may be expressed as the ratio of true or false positive or negative results respectively. These ratios should be determined at a number of different concentrations of the analyte, below, at and above the threshold concentration of the method. The result should be verified by using a separate reliable method, if available. If such an alternative reliable method is not available for comparison, a recovery test using suitable spiked and non-spiked samples of suitable matrices may be used.

$\% \text{ false positive} = (\text{false positive} \times 100) / \text{sum of the known negative}$

$\% \text{ false negative} = (\text{false negative} \times 100) / \text{sum of the known positive}$

6 Recovery

The problems with the use of recovery factors are outlined in the IUPAC Harmonised Guidelines For The Use Of Recovery Information In Analytical Measurement [ref].

7 Range

The range is the interval of concentration within which the analytical procedure demonstrates a suitable level of precision and accuracy.

Analysts should state the interval of concentration within which the analytical procedure demonstrates a suitable level of precision and accuracy.

It should usually be expressed with respect to the analyte or species in the test material presented for analysis, as the content range for which the analytical method may be used. It should be ensured that requirements on precision and trueness are met within the measurement range.

In situations where a statutory limit exists, this range may be defined, for example as the range from 0.75 to 1.25 times the legal norm for contents higher than or equal to 10^{-6} , and as the range corresponding to the norm \pm the norm * (2 * max.RSD_r/100) for contents below 10^{-6} .

For a zero tolerance the range may be defined as the limit of quantification + the limit of quantification * (2 * max. RSD_r/100).

The range of the method described in terms of the span of concentrations of analyte measurable in sample matrix, should correspond to a *working concentration range* of calibration standards. The range of the method is therefore dependent on the characteristics of the calibration and the detector response. The lower end of the content range corresponds to the lower limit for reliable quantification, defined as the limit of quantification. At the upper end of the range limitations will be imposed by various effects depending on the detection mechanism. In some cases it may be practical to use a non-linear standard curve in order to expand the working range, or on the basis of the use of a detector having an implicit non-linear response (for example a flame photometric sulfur detector in gas chromatography). Whether a linear or a non-linear working range is chosen, it is essential to establish the concentration range for which the applied standardisation relation is valid and also ensure that other quality requirements of the method are fulfilled.

8 Limit of Quantification

[Need to include reporting limit as example of administrative convenience]

The limit of quantification of an analytical procedure is the lowest amount or concentration of analyte in a sample which can be quantitatively determined with an acceptable level of precision and accuracy.

The limit of quantification should be stated if it is necessary to specify a lower limit of measurement below which acceptable accuracy and precision is not attained.

Using the method, carry out a number of independent determinations, preferably >20, using a sample which is known to contain the analyte at between 2 and 5 times the estimated detection limit. Calculate the relative standard deviation from the results.

$$\text{RSD } (s/x_{\text{mean}}) * 100\%$$

where: s = the standard deviation and

x_{mean} = the mean measured value.

It is necessary to decide what is or is not an acceptable relative level of accuracy for a particular application.

The limit of quantitation is the concentration at which an acceptable degree of performance, in terms of RSD%, is obtained. Although a figure of RSD = 10% is often quoted, a more useful guide to what is likely to be reasonable is the set of typical values of acceptable relative standard deviations for repeatability given earlier in this document.

It is usually the case that the limit of quantitation corresponds to the lowest standard concentration level in the calibration range.

(Note that the term *limit of determination* is sometimes used to correspond to a level based on 6 times the standard deviation of the concentration in a matrix blank - but this approach should be discouraged)

9 Limit of Detection

The limit of detection is the smallest amount or concentration of analyte in the test sample, that can be reliably distinguished, with stated significance, from the background or blank level.

It is less important to state the detection limit when evaluating methods for the determination of principal components of foods, than when carrying out analyses at trace level. Even in the latter situation, however, it may not always be considered necessary to quote a detection limit in cases where measurements are being made above the limit of quantitation, which is itself stated as a validated performance characteristic.

When measurements are being made above the limit of quantitation, and it is chosen to also quote a detection limit, a sufficient estimate of this detection limit can be made based on 3 times the standard deviation of the concentration in a matrix blank or one fortified with analyte close to this level. The number of determinations, which should preferably be at least 10, should be stated.

In cases where regulatory provisions require the absence or practical absence of certain analytes in food, in terms of a specified detection limit, it is recommended that a value for the limit of detection is calculated from a sum of the dispersion of blank readings and those for the dispersion of readings from a sample containing analyte at a level close to the estimated detection limit. It is recommended that this rigorous calculation is based on an assumption that all errors are random normal and that α and β are set at 0.05. According to this formulation, analyte concentrations which exceed the detection limit will have more than a 95% chance of detection, while blanks will have only a 5% chance of being mistakenly 'detected'.

$$\text{Limit of detection} = tS_b + tS_3$$

where

S_b = standard deviation calculated from blank determinations

S_s = standard deviation calculated from sample determinations at a level close to the blank

t = critical value for Student's 't' (one sided test):-
Significance level 0.05 (5%)

$$\text{Degrees of freedom} = 3$$

[Take IUPAC for the above]

For reliable calculation of the detection limit, a minimum of 10 (preferably >20) independent determinations should be carried out for the blank, to generate S_b , and a further 10 (preferably >20) independent determinations for the sample containing analyte at a level close to the blank, to generate S_s .

A matrix blank should be used in the above calculation. In cases where a matrix blank can not be obtained, an estimate of the background signal can be made at a representative part of the read-out, adjacent to the analyte signal, in the analyte-containing sample.

In some circumstances, it may be worthwhile to carry out an initial estimate of the detection limit. For an initial calculation not less than 8 analytical results are needed, 4 describing the blank variance and 4 results describing variance at a level close to the blank level.

In all cases when quoting a detection limit, the significance and reliability criteria, which includes the number and frequency (i.e. within-batch or between-batch) of determinations should be stated.

Note that, it is recommended that when quoting a detection limit, reliance is not placed on the estimation of detection limits using signal-to-noise ratios. In all cases when stating a limit of detection, if the recovery of the method is less than 100%, the limit of detection should be corrected for this recovery.

10 Sensitivity

The sensitivity of a method is a measure of the magnitude of the response caused by a certain amount of analyte.

Analysts should assess the ability of the method to discriminate between differences in analyte content.

Where the response has been established as linear with respect to concentration, sensitivity corresponds to the gradient of the response curve. It can also be represented by the angular coefficient of the standard curve in this case.

This term should not be used when referring to the detection capabilities of the measurement process.

11 Ruggedness

The ruggedness of an analytical method is the resistance to change of an analytical method when minor deviations are made in the experimental conditions of the procedure.

The aspects of the method which are susceptible to minor deviations in the experimental conditions of the procedure should be identified, and their influence on method performance evaluated, using ruggedness tests

The ruggedness of a method is tested by deliberately introducing small changes to the method and examining the effect on a particular aspect of performance. A large number of aspects of the method may need to be considered, but because most of these will have a negligible effect it will

normally be possible to vary several at once. An approach utilising 8 combinations of 7 variable factors has been described by Youden.

Some of the ruggedness parameters that a method should be evaluated for are:

Changes in instrument, operators, laboratories, brands or chemicals and supplies.

Slight changes in solution concentrations or sample preparation parameters e.g. pH, extraction time.

Changes in temperature

In addition, the stability of the standard and sample preparations must be assessed to determine the speed at which the analysis should be performed.

12 Fitness-for-Purpose

The ease of operation, in terms of sample throughput and costs, to achieve the required performance criteria and thereby meet the specified purpose.

The method selected for use should be practicable in terms of its ease of operation, for meeting the requirements of sample throughput and costs.

To validate a method in-house the criteria for acceptance, determined by the specification decided on prior to the validation study, must be met. In addition, account must be taken of information already available.

If for instance, the method which is chosen for use has already either been validated in a collaborative trial or alternatively been fully validated 'within laboratory', then it is only necessary to verify key performance parameters such as accuracy and precision, to ensure that the method is being implemented acceptably.